Reference 1

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 06-348755
(43)Date of publication of application : 22.12.1994

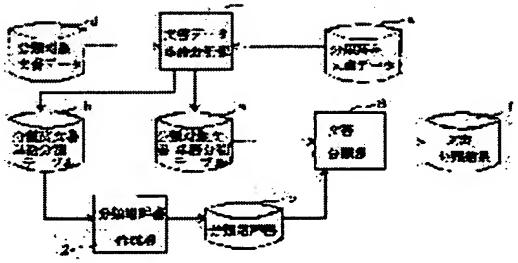
(51)Int.Cl. G06F 15/40

(21)Application number : 05-135588
(22)Date of filing : 07.06.1993

(71)Applicant : HITACHI LTD
(72)Inventor : KIYAMA TADAHIRO
TSUJI HIROSHI

(54) METHOD AND SYSTEM FOR CLASSIFYING DOCUMENT

(57)Abstract:
PURPOSE: To save huge human labor for classifying document data by preparing any specified dictionary for classification and classifying non-classified documents while utilizing this dictionary for classification.
CONSTITUTION: A document data word division part 1 divides the document data into words while referring to classified document data (a) and registers the classified result on a classified word division table (b). A dictionary for classification preparation part 2 detects the appearance frequency of words while referring to the classified word division table (b) and registers the result on a dictionary (c) for classification. After the dictionary (c) is generated, the word division part 1 divides the document data into words while referring to classification target document data (d) and registers the divided result on a classification target word division table (e). A document classification part 3 classifies the document data (e) while referring to the division table (e) and the dictionary (c) for classification and registers them on a document classified result (f). Thus, since the system is provided with the dictionary for classification preparing function and the document classifying function, the document data can be automatically classified.



LEGAL STATUS

[Date of request for examination] 18.05.2000
[Date of sending the examiner's decision of rejection] 07.01.2003
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
[Date of final disposal for application]
[Patent number]
[Date of registration]
[Number of appeal against examiner's decision of rejection]
[Date of requesting appeal against examiner's decision of rejection]
[Date of extinction of right]

Reference 1

(19)日本国特許庁（ＪＰ）　(12)**公開特許公報**（Ａ）　(11)特許出願公開番号

特開平6－348755

(43)公開日　平成６年(1994)12月22日

(51)Int.Cl.[5]	識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ０６Ｆ　15/40	５００　Ｔ	9194－5L		

審査請求　未請求　請求項の数12　ＯＬ　（全 20 頁）

(21)出願番号　特願平5－135588

(22)出願日　平成５年(1993)６月７日

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地

(72)発明者　木山　忠博
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内

(72)発明者　辻　洋
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内

(74)代理人　弁理士　小川　勝男

(54)【発明の名称】　文書分類方法およびそのシステム

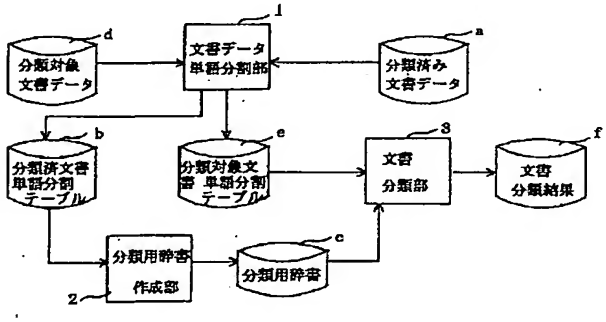
(57)【要約】

【目的】　人手による文書データの分類作業負担を軽減するために、複数の分類の文書データとを利用し、分類別のキーワードを抽出し分類用辞書を作成し、分類用辞書を利用して文書データを自動的に分類する方法およびシステムを提供することを目的とする。

【構成】　１は、ａを参照し、分類済文書データを単語分割し、ｂに登録する。また、１はｄを参照し、分類対象文書データを単語分割し、ｅに登録する。２は、ｂを参照し、分類別のキーワードを検出し、ｃに登録する。３は、ｅとｃを参照し、分類対象文書を分類し、ｆに登録する。

【効果】　従来は人手により分類されていた文書データを自動的に分類することが可能となり、人手による文書データの分類作業に費やす膨大な作業を省くことができるようになるという効果がある。

図１



【特許請求の範囲】
【請求項１】文書データを分類する処理システムにおいて、
一分類が一文書データ以上からなる分類済みの文書データから分類別のキーワードとなる語を抽出し分類用辞書を作成する手段と、
前記分類用辞書を用いて未分類の文書データを分類する手段を有する文書分類システム。
【請求項２】請求項１記載の文書分類システムにおいて、
分類用辞書を作成する手段は、
分類済みの文書データを用いて前記文書データ内の単語を検出し唯一の分類のみに出現する単語を検出し、前記単語を前記分類を表わすキーワードとして分類用辞書に登録することを特徴とする文書分類システム。
【請求項３】請求項１記載の文書分類システムにおいて、
分類用辞書を作成する手段は、
分類済みの文書データを用いて前記文書データを構成する項目を検出し項目別の単語を検出し唯一の分類のみに出現する単語を検出し、該単語は前記分類を表わし前記項目に出現するキーワードとして分類用辞書に登録することを特徴とする文書分類システム。
【請求項４】請求項１記載の文書分類システムにおいて、
分類用辞書を作成する手段は、
前記分類済み文書データを構成する項目が存在し利用者により前記分類用辞書を作成する範囲を前記項目により指定された場合に、前記分類済み文書データ中の指定された項目に相当する文書データのみを対象として前記分類用辞書を作成することを特徴とする文書分類システム。
【請求項５】請求項１記載の文書分類システムにおいて、
未分類の文書データを分類する手段は、
前記未分類文書データ中の単語を検出し前記分類用辞書に登録済みのキーワードとの一致数を検出し一致した分類の中で最も一致数が多い分類を前記未分類文書データの分類結果とすることを特徴とする文書分類システム。
【請求項６】請求項１記載の文書分類システムにおいて、
未分類の文書データを分類する手段は、
前記未分類文書データ中の単語を検出し前記分類用辞書に登録済みのキーワードとの一致数を検出し一致した分類の中で一致数が多い順番に優先度が高い分類として前記未分類文書データを分類することを特徴とする文書分類システム。
【請求項７】請求項１記載の文書分類システムにおいて、
未分類の文書データを分類する手段は、
前記未分類文書データを構成する項目が存在し利用者により前記未分類文書データを分類する範囲を前記項目により指定された場合に、前記未分類文書データ中の指定された項目に相当する文書データのみを処理の対象とすることを特徴とする文書分類システム。
【請求項８】文書データを分類する処理システムにおいて、
一分類が一文書データ以上からなる分類済みの文書データから分類別のキーワードとなる語を抽出し分類用辞書を作成する手段と、
前記分類用辞書を用いて未分類の文書データを分類する手段と、前記未分類文書データを分類する手段により分類された結果から新たにキーワードを検出し前記分類用辞書に登録する手段を有する文書分類システム。
【請求項９】請求項８記載の文書分類システムにおいて、
新たにキーワードを検出し前記分類用辞書に登録する手段は、
前記未分類文書データを分類する手段により分類された文書データ中の単語で前記分類用辞書中に存在しない単語であれば該単語は前記分類を表わすキーワードとして前記分類用辞書に登録することを特徴とする文書分類システム。
【請求項１０】請求項８記載の文書分類システムにおいて、
新たにキーワードを検出し前記分類用辞書に登録する手段は、
前記未分類文書データを分類する手段による分類結果が正しいか否か利用者に問合せ正しいと指示された場合に、前記未分類文書データを分類する手段により分類された文書データ中の単語で前記分類用辞書中に存在しない単語であれば該単語を前記分類を表わすキーワードとして前記分類用辞書に登録することを特徴とする文書分類システム。
【請求項１１】請求項８記載の文書分類システムにおいて、
新たにキーワードを検出し前記分類用辞書に登録する手段は、
前記未分類文書データを分類する手段による分類結果が正しいか否か利用者に問合せ正しいと指示された場合に、前記未分類文書データを分類する手段により分類された文書データ中の単語で前記分類用辞書中に存在しない単語であれば該単語を前記分類を表わすキーワードとして利用者に提示し、利用者が指示したキーワードのみを前記分類用辞書に登録することを特徴とする文書分類システム。
【請求項１２】請求項８記載の文書分類システムにおいて、
新たにキーワードを検出し前記分類用辞書に登録する手段は、

前記未分類文書データを分類する手段による分類結果が正しいか否か利用者に問合せ正しいと指示された場合に一致したキーワードを利用者に提示し、誤ったキーワードが存在している場合に利用者の指示により分類用辞書から該キーワードを削除することを特徴とする文書分類システム。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は文書データの自動分類方法に係り、特に、大量の文書データを自動的に分類する場合に好適な文書分類方法およびそのシステムに関する。

【０００２】

【従来の技術】従来の文書分類方法は、例えば、特開平２－１５８８７１号公報に記載されている。文書に含まれるキーワードの頻度値から各文書の概念特徴量を求め、これに応じて文書を分類している。

【０００３】

【発明が解決しようとする課題】従来の文書分類方法では、文書間の概念特徴量の差に応じて文書間の距離を求め、この距離によって文書の分類を行なっているが、概念特徴量を求めるために、文書分類用のシソーラスやキーワード分類項目を予め人手により登録しこれを利用している。つまり、文書を分類するために分類用の情報を人手により定義しなければならないという問題があった。

【０００４】本発明の目的は、上記問題を解決するために、複数に分類された書データと、分類内での各単語の出現頻度を利用し、分類別のキーワードを自動的に抽出し、分類用辞書を作成し、この分類用辞書を利用することにより未分類の文書を分類する方法およびシステムを提供することにある。

【０００５】

【課題を解決するための手段】上記目的を達成するために、本発明の文書分類方法は、一分類が一文書データ以上からなる分類済みの文書データから分類別のキーワードとなる語を抽出し分類用辞書を作成する方法と、分類用辞書を用いて未分類の文書データを分類する方法と、未分類文書データを分類する方法により分類された結果から新たにキーワードを検出し分類用辞書に登録する方法を具備する。また、本発明の文書分類システムは、上記文書分類方法を実行するプログラムを具備した自然言語処理システムである。

【０００６】

【作用】本発明の文書分類方法およびシステムは、まず、分類済みの文書データを取得する。次に、利用者により分類用辞書の作成範囲を文書データを構成する項目により指定されている場合は指定項目のみの文書データを処理対象とし、指定されてない場合は文書データ全体を処理対象とし分類用辞書を作成する。次に、分類済みの文書データから得た分類別に、唯一の分類のみに存在する単語を検出し、この単語を該当する分類のキーワードとして分類用辞書に登録する。また、文書データが複数の項目により構成されている場合にはキーワードと項目の対応関係も分類用辞書登録する。

【０００７】分類用辞書作成後、未分類の文書データを取得する。次に、利用者により文書データの分類範囲を文書データを構成する項目により指定されている場合は指定項目のみの文書データを処理対象とし、指定されてない場合は文書データ全体を処理対象とし文書を分類する。未分類文書データの単語を検出し、分類用辞書中のキーワードと比較照合し、分類別に一致回数を求める。

【０００８】次に、キーワードの一致回数が最も多い分類を分類対象文書データの分類結果とする。また、一致回数が多い分類の順番に優先度を付与し分類結果とする。また、未分類文書データを分類する方法による分類結果が正しいか否か利用者に問合せ正しいと指示された場合に、未分類文書データを分類する方法により分類された文書データ中の単語で分類用辞書中に存在しない単語であれば、この単語を該分類を表わすキーワードとして分類用辞書に登録する。また、キーワードを分類用辞書に登録する場合に、利用者が指示したキーワードのみを分類用辞書に登録する。

【０００９】また、未分類文書データを分類する方法による分類結果が正しいか否か利用者に問合せ正しいと指示された場合に一致したキーワードを利用者に提示し、誤ったキーワードが存在している場合に利用者の指示により分類用辞書から該キーワードを削除する。以上により、分類されてない文書データを自動的に分類することができると同時に、分類用辞書の自己増殖が可能となる。

【００１０】

【実施例】以下、本発明の実施例を、図面により詳細に説明する。図１は、本発明の文書分類方法の一実施例を示す機能ブロック図である。文書データ単語分割部１は、分類済文書データａを参照し、文書データを単語分割し、分割結果を分類済単語分割テーブルｂに登録する。分類用辞書作成部２は、分類済単語分割テーブルｂを参照し、単語の出現頻度を検出し、分類用辞書ｃに登録する。分類用辞書が生成された後、文書データ単語分割部１は、分類対象文書データｂを参照し、文書データを単語分割し、分割結果を分類対象単語分割テーブルｅに登録する。文書分類部３は、分類対象単語分割テーブルｅと分類用辞書を参照し、分類対象文書データｄを分類し文書分類結果ｆに登録する。

【００１１】このように分類用辞書作成機能及び文書分類機能を持たせることにより、文書データの自動分類を実現する。

【００１２】図から明らかなように、文書データ単語分割部１，分類用辞書作成部２，文書分類部３は処理を示し、分類済文書データａ，分類済単語分割テーブルｂ，

分類用辞書ｃ，分類対象文書データｄ，分類対象文書単語分割テーブルｅ，文書分類結果ｆはファイル（テーブルとも呼ぶ）である。このように、本実施例によれば、各機能ブロックが、プログラム論理により構成されている。そのため、各機能ブロック単位にＬＳＩ化が可能であり、文書分類装置として、処理の高速化を図ることができる。

【００１３】図２は、図１における文書分類装置の全体的なハードウェア構成を示すブロック図である。入出力装置４は、データの入力、および、各種情報の表示を行なう。プロセッサ５は、プログラムに基づき、図１における処理を実行する。記憶装置６は、図１における各種文書データａや各種プログラム等を格納する。さらに、記憶装置６は、プロセッサ５の各処理実行用のメモリであるワーキングエリアａｅ，文書データ単語分割部格納エリア１０，分類用辞書作成部格納エリア２０，文書分類部格納エリア３０，文書データ格納エリアａｄ，単語分割テーブル格納エリアｂｅ，分類用辞書格納エリアｃ，文書分類結果格納エリアｆの記憶部を持っている。

【００１４】記憶装置６に格納される各プログラムは、プロセッサ５において実行される。その実行に際して、必要に応じて入出力装置４が用いられる。

【００１５】図３は、図１における文書データ単語分割部の処理手順を表すＰＡＤ図（Problem Analysis Diagram）である。分類済文書データａから文書データを取得し単語分割を行ない分類済単語分割テーブルｂに格納するまでの処理、または、分類対象文書データｄから文書データを取得し単語分割を行ない分類対象単語分割テーブルｅに格納するまでの処理を示したものである。

【００１６】以下、この処理をＰＡＤ図に従って説明する。分類済文書データａを参照し、先頭文書データから末尾文書データまで以下の処理を行なう（ステップ１１）。まず、分類済文書データａから一文書分のデータを取得し（ステップ１２）、文書データを単語分割し見出し文字列，品詞をそれぞれ単語分割テーブルｂの見出し文字列ｂ１，品詞ｂ２に格納する（ステップ１３）。

【００１７】次に、文書データが項目で区分されているか判別し（ステップ１４）、項目で区分されている場合には、各単語に該当する項目を分類済単語分割テーブルｂの項目ｂ４に格納する（ステップ１５）。次に、該当する文書データの分野名，文書番号を分類済単語分割テーブルｂの分類名ｂ５，文書Ｎｏｂ６に格納する（ステップ１６）。次に、処理の対象を次の文書データに移動する（ステップ１７）。

【００１８】また、ステップ１１～ステップ１７において、分類対象文書データｄを入力とした場合には、分類対象単語分割テーブルｅに単語分割結果を格納する。ステップ１１～ステップ１７により、図４に示す分類済文書データａを取得し単語分割を行ない、図５に示す分類済単語分割テーブルｂに格納する。また、ステップ１１～ステップ１７により、図９に示す分類対象文書データｄを取得し単語分割を行ない、図１０に示す分類対象単語分割テーブルｅに格納する。

【００１９】図４は、分類済文書データａの例であり、分類名，文書番号付きの分類済みの文書データａの例である。分類済み文書データは「題名」「要旨」「目的」「主内容」「今後の課題」などの項目により構成されている。

【００２０】図５は、単語分割テーブルの例であり、見出し文字列ｂ１、品詞ｂ２、項目ｂ３、分類名ｂ４、文書Ｎｏｂ５の項目により構成されている。見出し文字列ｂ１は分割された単語を構成する文字列、品詞Ｃ２は単語の品詞、項目ｂ３は該当する単語に対応する項目の項目名、分類名ｂ４は該当する文書データの分類名、文書Ｎｏｂ５は文書データの識別番号を表している。

【００２１】図６は、図１における分類用辞書作成部２の処理を表わすＰＡＤ図であり、未分類の文書を分類するための辞書を作成する処理を示したものである。以下、この処理をＰＡＤ図に従って説明する。

【００２２】分類済単語分割テーブルｂを参照し、分類済単語分割テーブルの先頭レコードから末尾レコードまで以下の処理を行なう（ステップ２０１）。まず、分類済単語分割テーブルｂを参照し、１レコード分の情報を取得し（ステップ２０２）、見出し文字列ｂ１と品詞ｂ２と分類名Ｂ４が等しい他のレコードを分類済み単語分割テーブルｂより検索し保持しワークテーブルに格納する（ステップ２０３）。

【００２３】次に、該当する文書データを構成する項目が存在するか判別し（ステップ２０４）、項目が存在する場合に、項目別の各単語の出現回数を求めワークテーブルに格納する（ステップ２０５）。次に、文書データ全体の各単語の出現回数を求めワークテーブルに格納する（ステップ２０６）。

【００２４】次に、処理対象となるレコードを次のレコードへ移動する（ステップ２０７）。　ステップ２０１の繰返し処理が終了した後、２０１～２０６で格納したワークテーブルの先頭レコードから末尾レコードまで以下の処理を行なう（ステップ２０８）。

【００２５】まず、ワークテーブルより１レコード文の情報を取得し保持する（ステップ２０９）。次に、見出し文字列と品詞が等しい他のレコードがワークテーブル中に存在するか判別し（ステップ２１０）、存在しない場合に、保持した本レコードを分類用辞書３に格納する（ステップ２１１）。

【００２６】次に、処理対象となるレコードを次のレコードへ移動する（ステップ２１２）。

【００２７】ステップ２０１～ステップ２０７により、図５に示す分類済単語分割テーブルｂから図７に示すワークテーブルが生成され、ステップ２０８～ステップ２１２により、図７に示すワークテーブルから図８に示す

分類用辞書ｃが生成される。

【００２８】図７は、図６におけるワークテーブルの例であり、見出し文字列Ｗ１、品詞Ｗ２、題名Ｗ３～今後の課題Ｗ７、合計Ｗ８、分類名Ｗ９の項目により構成されている。見出し文字列Ｗ１は分割された単語を構成する文字列、品詞Ｗ２は単語の品詞、題名Ｗ３～今後の課題Ｗ７は文書データを構成する項目中に出現する単語の出現回数、合計Ｗ８は文書データ中に出現する単語の出現回数、分類名Ｗ９は該当する文書データの分類名を表している。

【００２９】図８は、図１における分類用辞書の例であり、見出し文字列Ｃ１、品詞Ｃ２、題名Ｃ３～今後の課題Ｃ７、合計Ｃ８、分類名Ｃ９の項目により構成されている。見出し文字列Ｃ１は分割された単語を構成する文字列、品詞Ｃ２は単語の品詞、題名Ｃ３～今後の課題Ｃ７は文書データを構成する項目中に出現する単語の出現回数、合計Ｃ８は文書データ中に出現する単語の出現回数、分類名Ｃ９は該当する文書データの分類名を表している。本分類用辞書は図７に示すワークテーブルとは異なり、見出し文字列が重複することはなく、一つの見出し文字列は一つの分類名に対応している。

【００３０】図９は、分類対象文書データａの例であり、文書番号付きの分類対象文書データｄの例である。分類対象文書データは「題名」「要旨」「目的」「主内容」「今後の課題」などの項目により構成されている。

【００３１】図１０は、分類対象文書単語分割テーブルｄの例であり、見出し文字列ｅ１、品詞ｅ２、項目ｅ３、分類名ｅ４、文書Ｎｏｅ５の項目により構成されている。見出し文字列ｅ１は分割された単語を構成する文字列、品詞ｅ２は単語の品詞、項目ｅ３は該当する項目の項目名、分類名ｅ４は該当する文書データの分類名、文書Ｎｏｅ５は文書データの識別番号を表している。

【００３２】図１１は、図１における文書分類部３の処理を表すＰＡＤ図である。本処理は、未分類の文書を分類用辞書を用いて分類する処理である。また、本処理の前提として、図１における分類対象文書データｄを文書データ単語分割部１により単語分割し、単語分割結果を分類対象文書単語分割テーブルｅに格納されているものとする。

【００３３】以下、この処理をＰＡＤ図に従って説明する。分類対象文書単語分割テーブルｂを参照し、本テーブルの先頭レコードから末尾レコードまで以下の処理を行なう（ステップ３０１）。まず、利用者が文書の分類の対象とする項目を指定しているか判別し（ステップ３０２）、指定していれば、該当する項目に対応するレコードに限定し一レコード分の情報を取得し（ステップ３０３）、指定してなければ、一レコード分の情報を取得する（ステップ３０４）。

【００３４】次に、取得したレコードを文字列変数ＭＩＤＡＳＨＩ１に格納する（ステップ３０５）。次に、分類用辞書ｃを参照し、分類用辞書の先頭レコードから末尾レコードまで以下の処理を行なう（ステップ３０６）。まず、利用者が文書の分類の対象とする項目を指定しているか判別し（ステップ３０７）、指定していれば、該当する項目に対応するレコードに限定し分類用辞書ｃから一レコード分の情報を取得し（ステップ３０８）、指定してなければ、一レコード分の情報を取得する（ステップ３０９）。

【００３５】次に、取得したレコードを文字列変数ＭＩＤＡＳＨＩ２に格納する（ステップ３１０）。次に、ＭＩＤＡＳＨＩ１とＭＩＤＡＳＨＩ２に格納した情報から見出し文字列と品詞を取得し、見出し文字列と品詞が一致するか判別し（ステップ３１１）、一致する場合、一致回数を分類名別にカウントしこれを保持し（ステップ３１２）、一致した見出し文字列と品詞を保持する（ステップ３１３）。

【００３６】次に、本繰返し処理の対象を分類用辞書の次のレコードへ移動する（ステップ３１４）。次に、ステップ３０６の繰返し処理が終了した後に、ステップ３０１の繰返し処理の対象を分類対象文書単語分割テーブルの次のレコードへ移動する（ステップ３１５）。次に、ステップ３０１の繰返し処理が終了した後に、上記で保持した一致回数が最も多い分類名を本文書の分類結果として文書分類結果ｆに格納し（ステップ３１６）、一致回数が１回以上の分類名を対象に分類名別に一致回数が多い順番に並べ換え、分類候補として文書分類結果ｆに格納し（ステップ３１７）、一致した見出し文字列と品詞と一致回数を文書分類結果ｆに格納する（ステップ３１８）。

【００３７】ステップ３０１～ステップ３１８により、図１０に示す分類対象文書単語分割テーブルｅと分類用辞書ｃをもとに文書を分類し図１２に示す文書分類結果ｆを得ることができる。この文書分類結果ｆを利用者に示すことにより分類結果の確認が可能となる。

【００３８】図１２は、図１における文書分類結果ｆの例である。本例には、３つの文書の分類結果を示している。１文書目は図９に示している文書Ｎｏ「１０１」の文書であり、結果として「言語処理」に分類されている。これは分類名「言語処理」が分類用辞書の単語の一致回数が最も高かったことを表わしている。また、一致した単語は「辞書」「自然語」「ＬＲ解析」等であり括弧内の数値は単語の一致回数を表わしている。また、「言語処理」の次に単語の一致回数が多かった分類候補が「通信」であることを表わしている。

【００３９】２文書目は図９に示している文書Ｎｏ「１０２」の文書であり、結果として「電気回路」に分類されている。これは分類名「電気回路」が分類用辞書の単語の一致回数が最も高かったことを表わしている。また、一致した単語は「ＡＤ変換」「アナログ」「ディジタル」等であり括弧内の数値は単語の一致回数を表わし

ている。また、「電気回路」の次に単語の一致回数が多かった分類候補は無かったことを表わしている。

【００４０】３文書目は図９に示している文書Ｎｏ「１０３」の文書であり、結果として「通信」に分類されている。これは分類名「通信」が分類用辞書の単語の一致回数が最も高かったことを表わしている。また、一致した単語は「ネットワーク」「プロトコル」「ＬＡＮ」等であり括弧内の数値は単語の一致回数を表わしている。また、「通信」の次に単語の一致回数が多かった分類候補が「電気回路」であることを表わしている。

【００４１】以上、述べたように、本実施例によれば、従来は人手により分類されていた文書データを自動的に分類することが可能となり、人手による文書データの分類作業に費やす膨大な作業を省くことができるようになるという効果がある。また、パソコンやユニックスなどのニュースサービス（掲示板）は文書の内容によって投稿すべきニュースグループが非常に多く、どのニュースグループに投稿すべきであるか判断できないことがあるが、本実施例によれば、ニュースグループにより分類されている文書から分類用辞書を作成し自動分類可能となり、投稿対象文書のニュースグループを指定しなくとも自動的に投稿すべきニュースグループを知ることができる。また、自動的に分類し投稿することができる。

【００４２】次に、本発明の文書分類方法の拡張例について説明する。図１３は、本発明の文書分類方法において分類した結果を利用者に確認し、確認した結果から新たにキーワードとして分類用辞書に登録することを実現する機能ブロック図である。文書分類確認部４は文書分類結果ｆを参照し、分類した文書の分類結果や一致したキーワードを利用者に提示し、利用者の確認を促し、一致したキーワードで相応しく無いものがあれば、そのキーワードを利用者の指定により分類用辞書から削除し、分類した文書に新たなキーワードが存在するか判別し、存在する場合には新たなキーワードを分類用辞書に登録し、分類用辞書の自己増殖を行なう。

【００４３】図１４の［画面１］は、分類対象文書の分類結果の確認例である。提示内容として、文書番号、題名、分類結果、分類候補がある。また、分類結果と同時に、分類用辞書中のキーワードと一致したキーワードとキーワードの一致回数を示している。これらの情報から利用者が分類結果が正しいか否か判断し、分類結果の正否を入力する。

【００４４】図１４の［画面２］は、［画面１］において分類結果が正しいと利用者が判断した後、分類の根拠となった分類用辞書中の該分類のキーワードと一致したキーワードを利用者に提示し、キーワードとして相応しく無いものがあれば、利用者に指定するよう促している例である。また、指定された場合には、そのキーワードを分類用辞書から削除し、分類用辞書の精度を向上させる。本例では、そのようなキーワードは無いと利用者が応答している例を表わしている。

【００４５】図１５の［画面３］は、分類結果が正しかった場合に、分類用辞書には存在せず、分類対象文書にのみにキーワードが存在する場合に、このキーワードを該分類の新たなキーワードとして自動登録するか否か、また、キーワードを確認した後に、利用者がキーワードを種々選択し登録するか否か利用者の応答を促している例である。本例では、新たに出現したキーワードは確認せず自動登録するよう利用者が応答している例を表わしている。

【００４６】以上、述べたように、本拡張例によれば、分類用辞書中の任意の分類として誤ったキーワードが登録されている場合に、分類対象文書と一致したキーワードを利用者が確認し、誤ったキーワードを削除することにより、分類用辞書の精度向上を容易に実現することができる。また、分類対象文書に新たなキーワードが出現している場合に、そのキーワードを分類用辞書に自動登録する、または、利用者に提示し必要なキーワードのみ分類用辞書に登録することにより、利用者が時間をかけて辞書登録することなく、簡単に分類用辞書へのキーワードの辞書登録または自己増殖が可能となる。

【００４７】次に、本発明の文書分類方法の別の拡張例について説明する。図１６は、分類用辞書を作成する場合、または、分類対象文書を分類する場合に、文書データが一定の項目により区分されているならば、文書データ全体を分類用辞書作成または分類対象にするのではなく、文書内の項目を利用者が指定することを可能とし、その指定された項目のみを処理の対象とすることを実現する機能ブロック図である。文書分類対象項目指定部６は利用者が処理対象として指定した項目を認識し、分類用辞書を作成する場合には分類済文書データａを参照し文書データから該当する項目の文書データを取得し、文書データを分類する場合には分類対象文書データｄを参照し文書データから該当する項目の文書データを取得し、以降の処理で取得した文書データのみを処理対象とする。

【００４８】また、単に項目を指定するのではなく、任意の項目間の論理関係の指定及び認識を可能とすることにより、例えば、項目「主内容」または項目「今後の課題」に出現する単語、項目「主内容」と項目「今後の課題」の両項目に出現する単語を処理対象とすることが容易に実現可能である。

【００４９】以上、述べたように、本拡張例によれば、文書データ全体を処理対象とせず項目を限定することから、文書データの性格によっては分類用辞書精度向上や分類精度向上が期待できると同時に、膨大な量の文書データや文書データ自体が大きい場合にも実用的な処理速度を維持することが可能となり、処理速度が向上するという効果がある。

【００５０】

【発明の効果】本発明によれば、従来は人手により分類されていた文書データを自動的に分類することが可能となり、人手による文書データの分類作業に費やす膨大な作業を省くことができるようになるという効果がある。また、パソコンやユニックスなどのニュースサービス（掲示板）は文書の内容によって投稿すべきニュースグループが非常に多く、どのニュースグループに投稿すべきであるか判断できないことがあるが、本実施例によれば、ニュースグループにより分類されている文書から分類用辞書を作成し自動分類可能となり、投稿対象文書のニュースグループを指定しなくとも自動的に投稿すべきニュースグループを知ることができる。また、自動的に分類し投稿することができる。また、分類用辞書へのキーワード登録が簡単に行なうことができると同時に、分類用辞書の自己増殖も可能となるという効果がある。

【図面の簡単な説明】

【図１】本発明を施した文書分類方法の一実施例を示す機能ブロック図。

【図２】図１における文書分類方法の全体的なハードウェア構成を示すブロック図。

【図３】図１における文書データ単語分割プログラムのＰＡＤ図。

【図４】図１における分類済文書データの例。

【図５】図１における分類済単語分割テーブルの例。

【図６】図１における分類用辞書作成プログラムのＰＡＤ図。

【図７】図１における分類用辞書を作成するためのワークテーブルの例。

【図８】図１における分類用辞書の例。

【図９】図１における分類対象文書データの例。

【図１０】図１における分類対象文書単語分割テーブルの例。

【図１１】図１における文書分類プログラムのＰＡＤ図。

【図１２】図１における文書分類結果例。

【図１３】文書分類結果の確認方法及び分類用辞書学習方法を示す機能ブロック図。

【図１４】図１３における分類結果確認及び分類用辞書学習における利用者との対話例。

【図１５】図１３における分類結果確認及び分類用辞書学習における利用者との対話例。

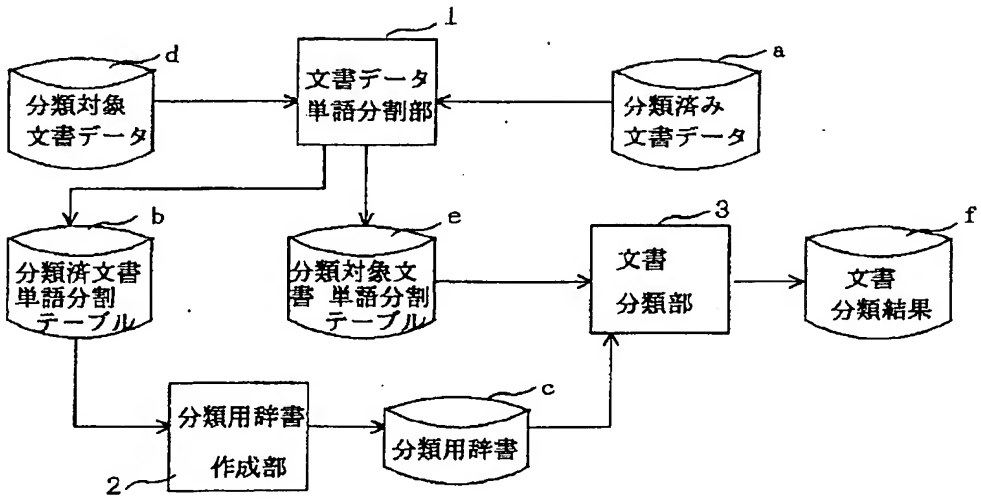
【図１６】文書分類方法において処理対象とする文書内の項目指定方法を示す機能ブロック図である。

【符号の説明】

１…文書データ単語分割部、２…分類用辞書作成部、３…文書分類部、ａ…分類済文書データ、ｂ…分類済単語分割テーブル、ｃ…分類用辞書、ｄ…分類対象文書データ、ｅ…分類対象文書単語分割テーブル。

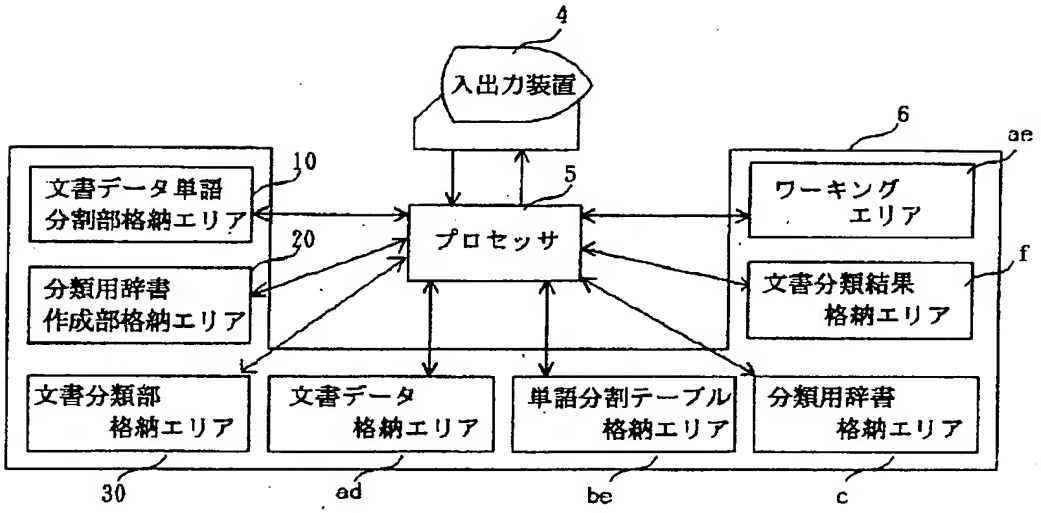
【図１】

図１



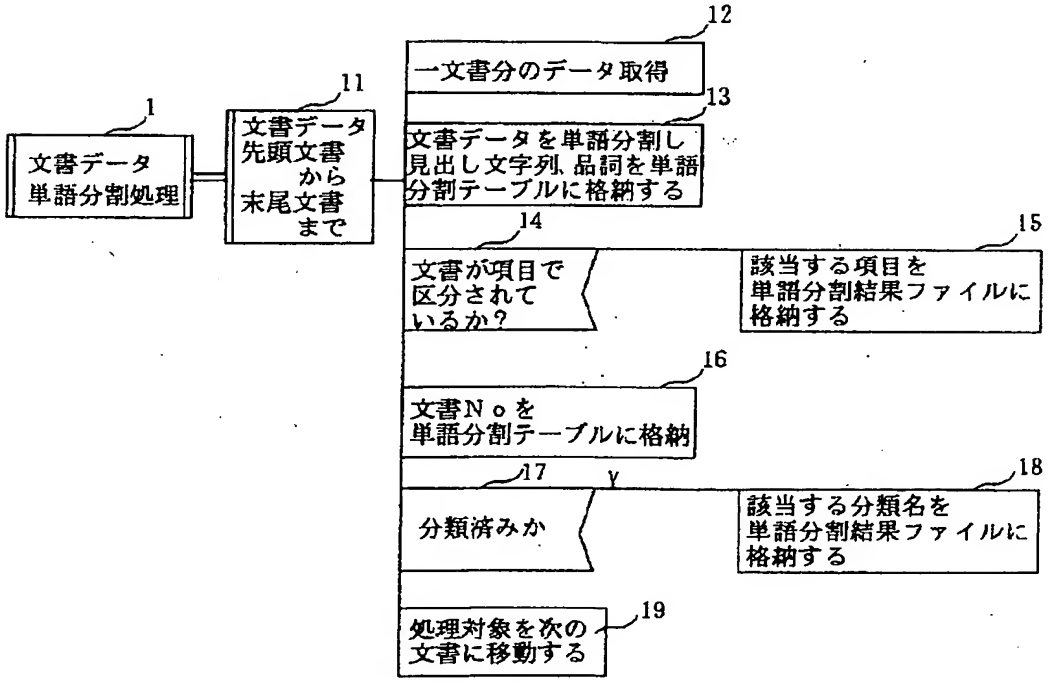
【図２】

図２


4
入出力装置
6
ae
10
文書データ単語
分割部格納エリア
5
プロセッサ
ワーキング
エリア
20
分類用辞書
作成部格納エリア
f
文書分類結果
格納エリア
文書分類部
格納エリア
文書データ
格納エリア
単語分割テーブル
格納エリア
分類用辞書
格納エリア
30
ad
be
c

【図３】

図３


1
文書データ
単語分割処理
11
文書データ
先頭文書
から
末尾文書
まで
12
一文書分のデータ取得
13
文書データを単語分割し
見出し文字列、品詞を単語
分割テーブルに格納する
14
文書が項目で
区分されて
いるか？
15
該当する項目を
単語分割結果ファイルに
格納する
16
文書Ｎｏを
単語分割テーブルに格納
17
y
分類済みか
18
該当する分類名を
単語分割結果ファイルに
格納する
19
処理対象を次の
文書に移動する

【図４】

図４

｛分類名｝：言語処理
｛文書Ｎｏ｝：１
＜題名＞
辞書生成方法に関する一考察
＜要旨＞
自然語インタフェースが参照する辞書情報を生成する処理システムにおいて、ＤＢ内の単語を構成する文字列を辞書情報の見出し文字列とする単語情報を生成し、見出し文字列とＤＢ内の単語との対応を持つ辞書の生成方式に関する考察結果について述べる
＜目的＞・・・・・・・・・・・・・・・・・・・・・・
・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・
＜主内容＞
・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・
＜今後の課題＞
・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・

｛分類名｝：電気回路
｛文書Ｎｏ｝：２
＜題名＞
ＡＤ変換の拡張
＜要旨＞
アナログからディジタルへ・・・・・・・・・・・・・・・・・・・
・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・
＜目的＞・・・・・・・・・・・・・・・・・・・・・・
・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・
＜主内容＞
・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・
＜今後の課題＞
・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・

｛分類名｝：通信
｛文書Ｎｏ｝：３
＜題名＞
パソコンＬＡＮの実例
＜要旨＞
高度情報通信時代を向かえ・・・・・・・・・・・・・・・・・・・
・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・
＜目的＞・・・・・・・・・・・・・・・・・・・・・・
・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・
＜主内容＞
・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・
＜今後の課題＞
・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・

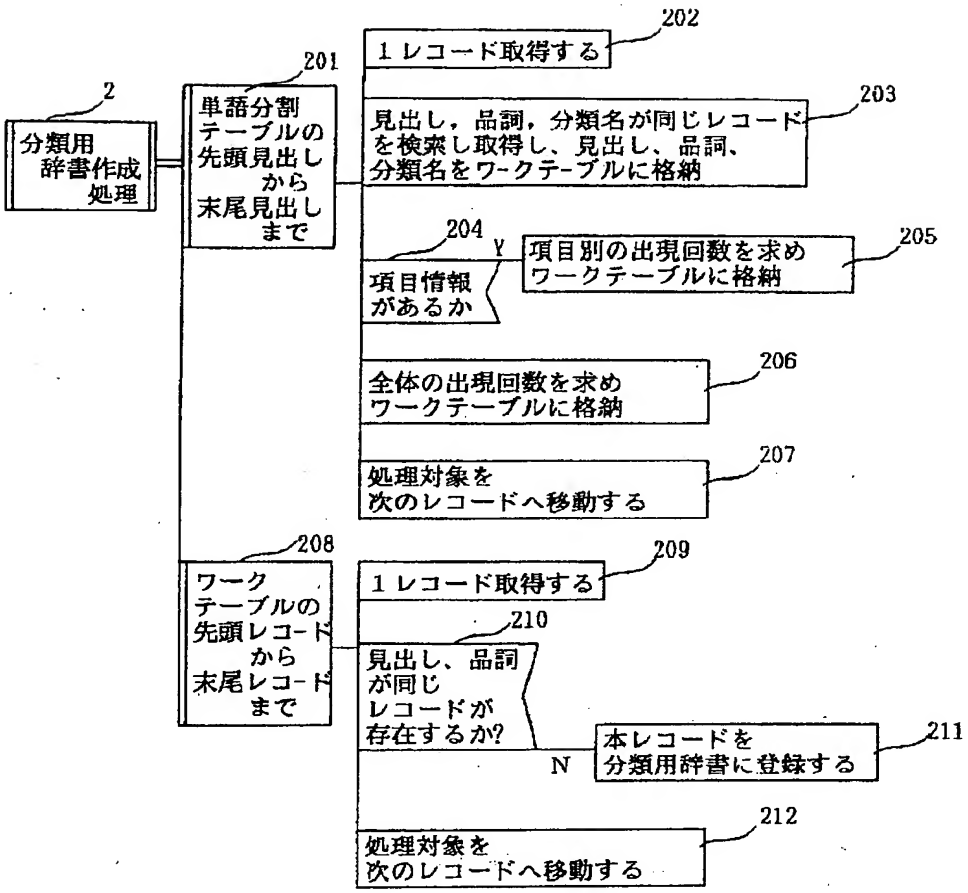
【図5】

図5

b1 見出し文字列	b2 品詞	b3 項目	b4 分類名	b5 文書No
題名	名詞	題名	言語処理	1
辞書	名詞	題名	言語処理	1
生成	名詞	題名	言語処理	1
方法	名詞	題名	言語処理	1
関	動詞	題名	言語処理	1
一	名詞	題名	言語処理	1
考察	サ変動詞	要旨	言語処理	1
要旨	名詞	要旨	言語処理	1
自然語	名詞	要旨	言語処理	1
インタフェース	名詞	要旨	言語処理	1
参照	サ変動詞	要旨	言語処理	1
・	・	・	・	・
・	・	・	・	・
・	・	・	・	・
ＡＤ変換	名詞	題名	電気回路	2
拡張	サ変名詞	題名	電気回路	2
アナログ	名詞	要旨	電気回路	2
ディジタル	名詞	要旨	電気回路	2
周波数	名詞	主内容	電気回路	2
・	・	・	・	・
・	・	・	・	・
・	・	・	・	・
パソコン	名詞	題名	通信	3
ＬＡＮ	動詞	題名	通信	3
実例	名詞	題名	通信	3
・	・	・	・	・
・	・	・	・	・
・	・	・	・	・

【図６】

図６


2
分類用辞書作成処理
201
単語分割テーブルの先頭見出しから末尾見出しまで
202
１レコード取得する
203
見出し，品詞，分類名が同じレコードを検索し取得し、見出し、品詞、分類名をワークテーブルに格納
204
項目情報があるか
Y
205
項目別の出現回数を求めワークテーブルに格納
206
全体の出現回数を求めワークテーブルに格納
207
処理対象を次のレコードへ移動する
208
ワークテーブルの先頭レコードから末尾レコードまで
209
１レコード取得する
210
見出し、品詞が同じレコードが存在するか?
N
211
本レコードを分類用辞書に登録する
212
処理対象を次のレコードへ移動する

【図７】

図７

W1	W2	W3	W4	W5	W6	W7	W8	W9
見出し文字列	品　詞	題名	要旨	目的	主内容	今後の課題	合　計	分類名
自然語	名詞	0	1	1	13	2	17	言語処理
辞書	名詞	1	2	1	15	2	21	言語処理
生成	名詞	1	3	1	21	3	29	言語処理
方法	名詞	1	0	1	20	2	24	言語処理
結果	名詞	0	1	1	7	3	12	言語処理
述べ	動詞	0	1	0	10	2	13	言語処理
パソコン	名詞	0	0	0	1	1	2	言語処理
ＬＲ解析	名詞	0	0	0	2	0	2	言語処理
・	・	・	・	・	・	・	・	・
・	・	・	・	・	・	・	・	・
・	・	・	・	・	・	・	・	・
ＡＤ変換	名詞	1	1	1	6	2	11	電気回路
拡張	サ変動詞	1	1	1	2	3	8	電気回路
アナログ	動詞	0	2	1	12	2	17	電気回路
ディジタル	名詞	0	1	1	13	1	16	電気回路
述べ	動詞	0	1	0	5	1	7	電気回路
実例	名詞	0	0	0	1	0	1	電気回路
生成	サ変動詞	0	0	0	2	0	2	電気回路
結果	名詞	0	0	0	3	1	4	電気回路
周波数	名詞	0	0	0	2	1	3	電気回路
・	・	・	・	・	・	・	・	・
・	・	・	・	・	・	・	・	・
・	・	・	・	・	・	・	・	・
パソコン	名詞	1	1	0	3	1	6	通信
ＬＡＮ	名詞	1	1	1	4	2	9	通信
ネットワーク	名詞	0	1	1	5	1	8	通信
プロトコル	名詞	0	0	0	2	1	3	通信
実例	名詞	1	1	1	2	1	6	通信
拡張	サ変動詞	0	0	0	1	1	2	通信
・	・	・	・	・	・	・	・	・
・	・	・	・	・	・	・	・	・
・	・	・	・	・	・	・	・	・

【図８】

図８

C1	C2	C3	C4	C5	C6	C7	C8	C9
見出し文字列	品　詞	題名	要旨	目的	主内容	残の課題	合　計	分類名
自然語	名詞	0	1	1	13	2	17	言語処理
辞書	名詞	1	2	1	15	2	21	言語処理
ＬＲ解析	名詞	0	0	0	3	1	4	言語処理
・	・	・	・	・	・	・	・	・
・	・	・	・	・	・	・	・	・
・	・	・	・	・	・	・	・	・
ＡＤ変換	名詞	1	1	1	6	2	11	電気回路
アナログ	動詞	0	2	1	12	2	17	電気回路
ディジタル	名詞	0	1	1	13	1	16	電気回路
周波数	名詞	0	0	0	2	1	3	電気回路
・	・	・	・	・	・	・	・	・
・	・	・	・	・	・	・	・	・
・	・	・	・	・	・	・	・	・
ＬＡＮ	名詞	1	1	1	4	2	9	通信
ネットワーク	名詞	0	1	1	5	1	8	通信
プロトコル	名詞	0	0	0	2	1	3	通信
・	・	・	・	・	・	・	・	・
・	・	・	・	・	・	・	・	・
・	・	・	・	・	・	・	・	・

【図9】

図9

{文書No} ：101
<題名>
機械翻訳用辞書構ツールの開発
<要旨>
自然語処理技術の研究において、機械翻訳用の辞書を構築するためのツールの開発を行なった。
<目的>
<主内容>
<今後の課題>

{文書No} ：102
<題名>
AD変換の拡張（その2）
<要旨>
前回のアナログからディジタルへ
<目的>
<主内容>
<今後の課題>

{文書No} ：103
<題名>
ネットワークを用いたデータベース検索
<要旨>
ネットワークを用いたデータベース検索システムを開発した。本報ではプロトコルの
<目的>
<主内容>
<今後の課題>

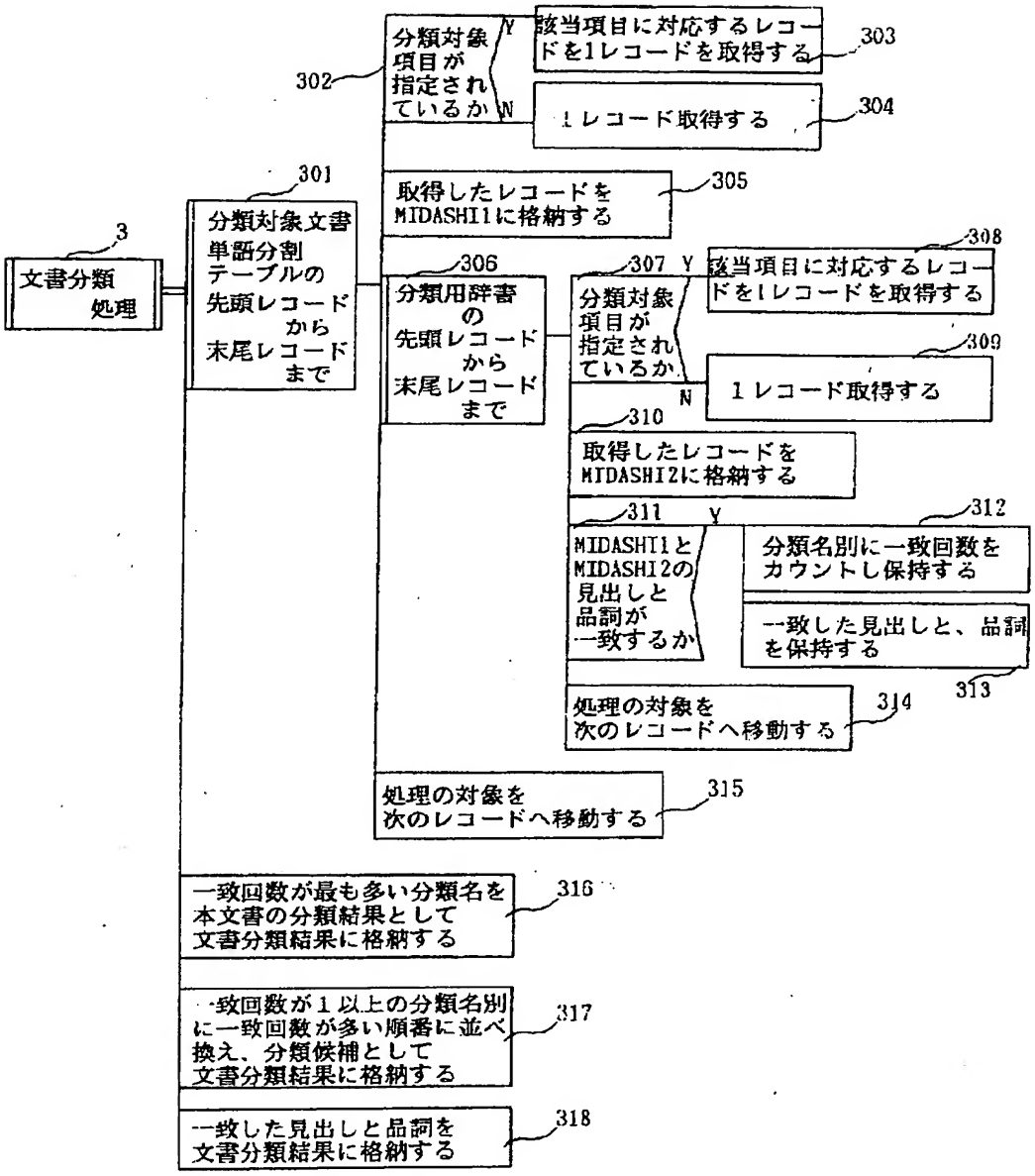
【図１０】

図１０

e1 見出し文字列	e2 品　詞	e3 項　目	e4 分類名	e5 文書No
題名	名詞	題名	なし	１０１
機械翻訳	名詞	題名	なし	１０１
用	名詞	題名	なし	１０１
辞書	名詞	題名	なし	１０１
構築	サ変動詞	題名	なし	１０１
ツール	名詞	題名	なし	１０１
開発	サ変動詞	要旨	なし	１０１
要旨	名詞	要旨	なし	１０１
自然語	名詞	要旨	なし	１０１
処理	サ変動詞	要旨	なし	１０１
技術	名詞	要旨	なし	１０１
ＬＲ解析	名詞	要旨		１０１
・	・	・	・	・
・	・	・	・	・
・	・	・	・	・
ＡＤ変換	名詞	題名	なし	１０２
拡張	サ変名詞	題名	なし	１０２
その	代名詞	題名	なし	１０２
２	数字	題名	なし	１０２
前回	名詞	要旨	なし	１０２
アナログ	名詞	要旨	なし	１０２
ディジタル	名詞	要旨	なし	１０２
・	・	・	・	・
・	・	・	・	・
・	・	・	・	・
ネットワーク	名詞	題名	なし	１０３
用い	動詞	題名	なし	１０３
データベース	名詞	題名	なし	１０３
検索	名詞	題名	なし	１０３
プロトコル	名詞	要旨	なし	１０３
周波数	名詞	主内容	なし	１０３
・	・	・	・	・
・	・	・	・	・
・	・	・	・	・

【図１１】

図１１


3
文書分類処理
301
分類対象文書単語分割テーブルの先頭レコードから末尾レコードまで
302
分類対象項目が指定されているか
Y
該当項目に対応するレコードを1レコードを取得する
303
N
1レコード取得する
304
取得したレコードをMIDASHI1に格納する
305
306
分類用辞書の先頭レコードから末尾レコードまで
307
分類対象項目が指定されているか
Y
該当項目に対応するレコードを1レコードを取得する
308
N
1レコード取得する
309
取得したレコードをMIDASHI2に格納する
310
311
MIDASHI1とMIDASHI2の見出しと品詞が一致するか
Y
分類名別に一致回数をカウントし保持する
312
一致した見出しと、品詞を保持する
313
処理の対象を次のレコードへ移動する
314
処理の対象を次のレコードへ移動する
315
一致回数が最も多い分類名を本文書の分類結果として文書分類結果に格納する
316
一致回数が１以上の分類名別に一致回数が多い順番に並べ換え、分類候補として文書分類結果に格納する
317
一致した見出しと品詞を文書分類結果に格納する
318

【図１２】

図１２

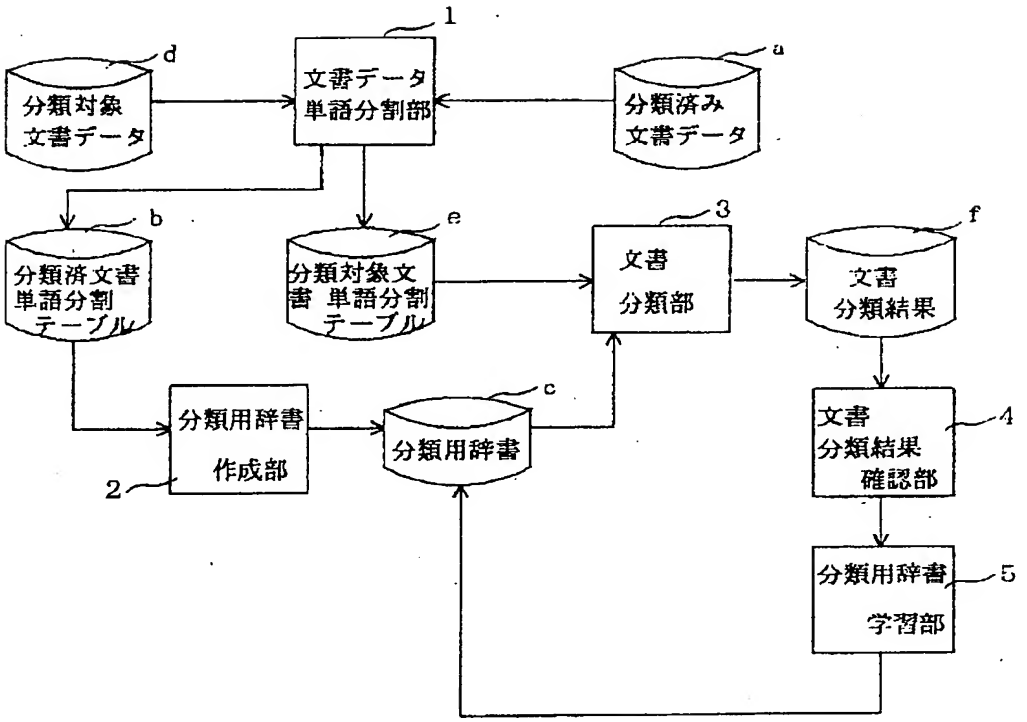
[文書Ｎｏ]　：　１０１
[題名]　：　機械翻訳用辞書構築ツールの開発
[分類結果]　：　言語処理（辞書(25),自然語(23),ＬＲ解析(6)･･･）
[分類候補]　：　通信（ネットワーク(3)）

[文書Ｎｏ]　：　１０２
[題名]　：　ＡＤ変換の拡張（その２）
[分類結果]　：　電気回路（ＡＤ変換(35),アナログ(31),ディジタル(30)･･･）
[分類候補]　：　なし

[文書Ｎｏ]　：　１０３
[題名]　：　ネットワークを用いたデータベース検索
[分類結果]　：　通信（ネットワーク(40),プロトコル(25),Ｌ.ＡＮ(11)･･･）
[分類候補]　：　電気回路（周波数(2)）

【図１３】

図１３



【図１５】

図１５

［画面３］

登録しない　確認後登録　自動登録

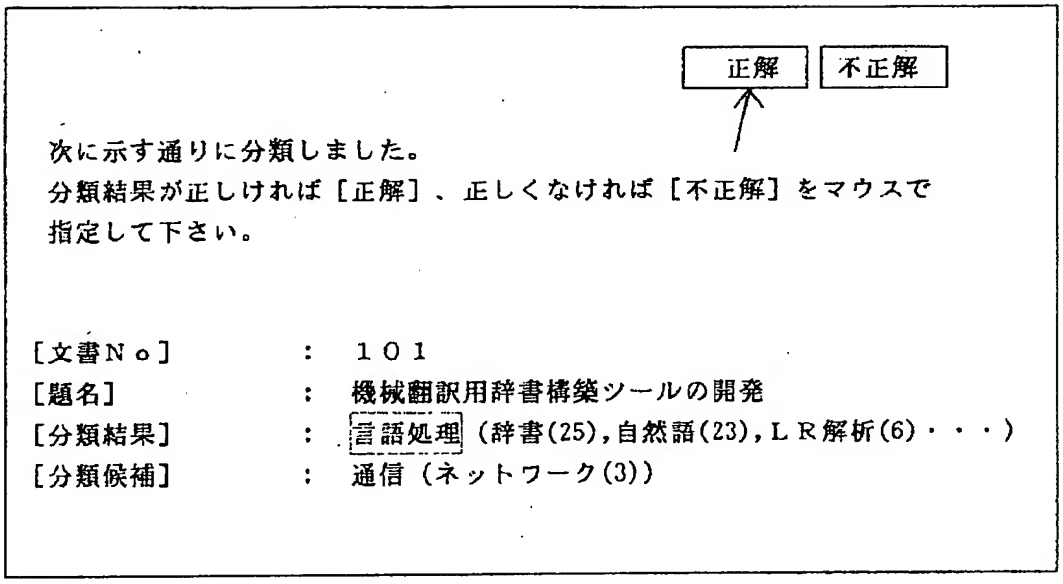
本分類結果をもとに、分類用辞書に新たにキーワードを登録できます。
キーワードの登録方法として３通りあります。
該当する登録方法をマウスで指定して下さい。

［文書Ｎｏ］　：　１０１
［題名］　：　機械翻訳用辞書構築ツールの開発
［分類結果］　：　言語処理（辞書(25)，自然語(23)，ＬＲ解析(6)・・・）
［新キーワード］　：　構文解析，形態素，統語

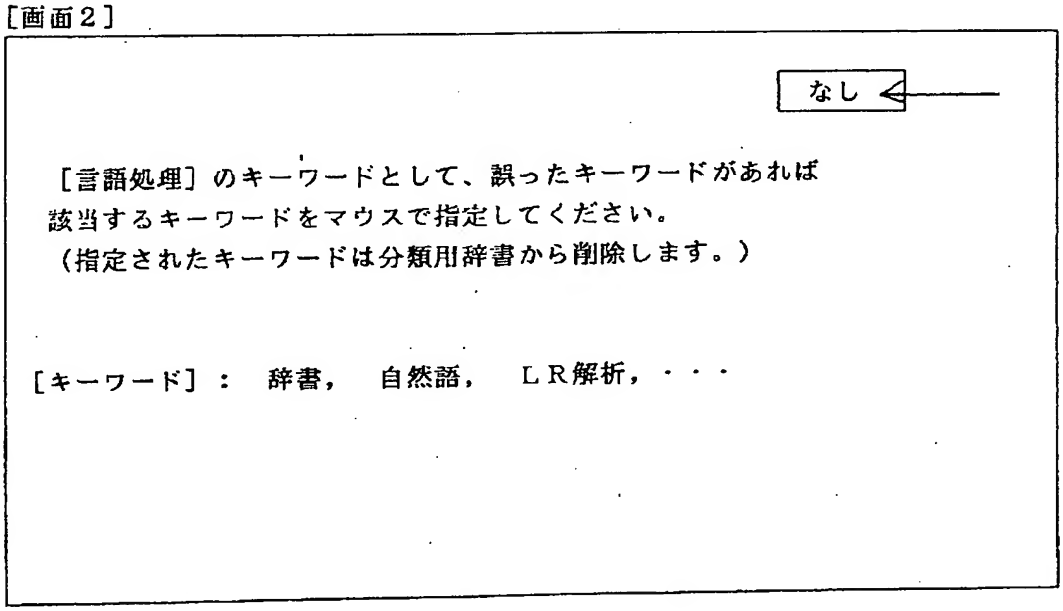
【図１４】

図１４

[画面１]

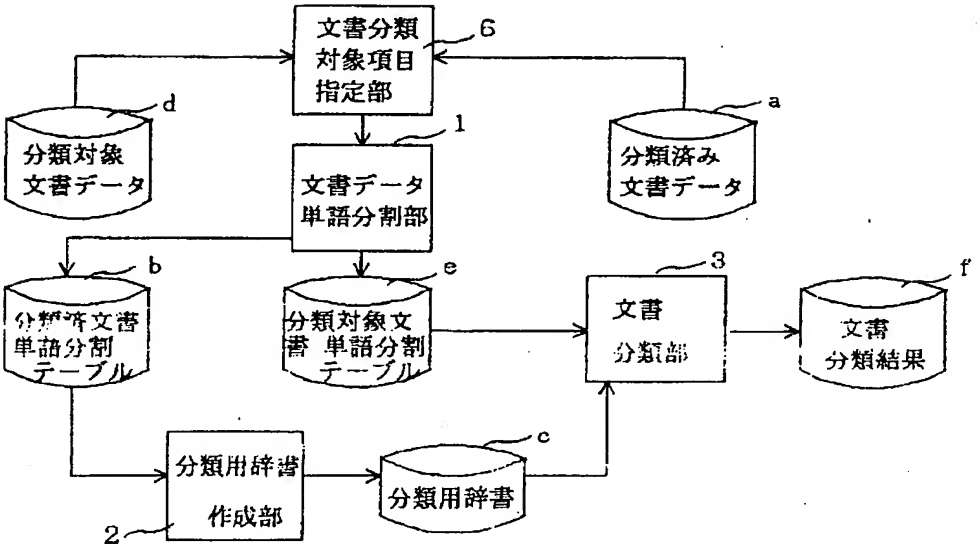


[画面２]



【図１６】

図１６


文書分類
対象項目
指定部
6
d
分類対象
文書データ
1
文書データ
単語分割部
a
分類済み
文書データ
b
分類済文書
単語分割
テーブル
e
分類対象文
書 単語分割
テーブル
3
文書
分類部
f
文書
分類結果
分類用辞書
作成部
2
c
分類用辞書

【公報種別】特許法第１７条の２の規定による補正の掲載
【部門区分】第６部門第３区分
【発行日】平成１３年２月９日（２００１．２．９）

【公開番号】特開平６－３４８７５５
【公開日】平成６年１２月２２日（１９９４．１２．２２）
【年通号数】公開特許公報６－３４８８
【出願番号】特願平５－１３５５８８
【国際特許分類第７版】
　G06F　15/40　　500
【ＦＩ】
　G06F　15/40　　500

【手続補正書】
【提出日】平成１２年５月１８日（２０００．５．１８）
【手続補正１】
【補正対象書類名】明細書
【補正対象項目名】特許請求の範囲
【補正方法】変更
【補正内容】
【特許請求の範囲】
【請求項１】文書データを分類する方法において、
１つの分類が少なくとも１つの文書データからなる分類済みの文書データから分類別のキーワードとなる語抽出して分類用辞書を作成し、
前記分類用辞書を用いて未分類の文書データを分類することを特徴とする文書分類方法。
【請求項２】文書データを分類する処理システムにおいて、
１つの分類が少なくとも１つの文書データからなる分類済みの文書データから分類別のキーワードとなる語抽出して分類用辞書を作成する手段、
前記分類用辞書を用いて未分類の文書データを分類する手段を有することを特徴とする文書分類システム。
【請求項３】前記分類用辞書を作成する手段は、
分類済みの文書データを用いて前記文書データ内の単語を検出して唯一の分類に出現する単語を検出し、前記単語を前記分類を表わすキーワードとして前記分類用辞書に登録することを特徴とする請求項２記載の文書分類システム。
【請求項４】前記未分類の文書データを分類する手段は、
前記未分類文書データ中の単語を検出して前記分類用辞書に登録済みのキーワードとの一致数を検出し、一致した分類の中で最も一致数が多い分類を前記未分類文書データの分類結果とすることを特徴とする請求項２記載の文書分類システム。